Extracting a Knowledge Base of Mechanisms from COVID-19 Papers

Tom Hope[♣,♠*] **Aida Amini**[♣*] **David Wadden**[♣]
Madeleine van Zuylen[♠] **Sravanthi Parasa**[♡] **Eric Horvitz**[◇]
Daniel Weld[♣,♠] **Roy Schwartz**[♥] **Hannaneh Hajishirzi**[♣,♠]
[♣] Paul G. Allen School for Computer Science & Engineering, University of Washington
[♠] Allen Institute for Artificial Intelligence [♡] Swedish Medical Group
[◇] Microsoft Research [♥] The Hebrew University of Jerusalem, Israel
{tomh,aidaa,danw, hannah}@allenai.org

Abstract

The COVID-19 pandemic has spawned a diverse body of scientific literature that is challenging to navigate, stimulating interest in automated tools to help find useful knowledge. We pursue the construction of a knowledge base (KB) of *mechanisms*—a fundamental concept across the sciences, which encompasses activities, functions and causal relations, ranging from cellular processes to economic impacts. We extract this information from the natural language of scientific papers by developing a broad, unified schema that strikes a balance between relevance and breadth. We annotate a dataset of mechanisms with our schema and train a model to extract mechanism relations from papers. Our experiments demonstrate the utility of our KB in supporting interdisciplinary scientific search over COVID-19 literature, outperforming the prominent PubMed search in a study with clinical experts. Our search engine, dataset and code are publicly available.[1]

1 Introduction

> "Some experts are familiar with one field, such as AI or nanotechnology [...] no one is capable of connecting the dots and seeing how breakthroughs in AI might impact nanotechnology, or vice versa."
> –*Yuval Noah Harari, Homo Deus, 2016*

The effort to mitigate the COVID-19 pandemic is an interdisciplinary endeavor the world has rarely seen (Apuzzo and Kirkpatrick, 2020). As one recent example, expertise in virology, physics, epidemiology and engineering enabled a group of 200 scientists to understand and bring attention to the airborne transmissibility of the SARS-CoV-2 virus (Morawska et al., 2020). The diverse and rapidly expanding body of past and present findings related to COVID-19 (Wang et al., 2020b) makes it challenging to keep up, hindering scientists' pace in making new discoveries and connections.

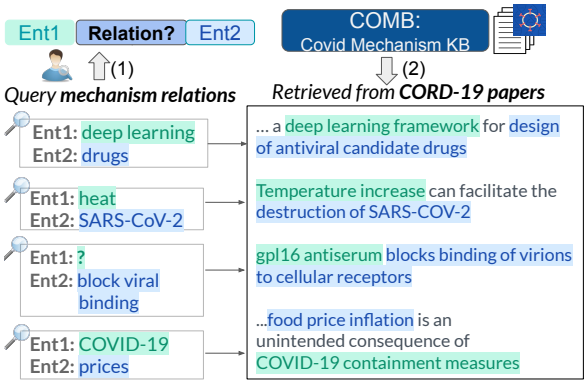


Figure 1: Our COVID-19 Mechanism KB (COMB) is extracted from scientific papers and can be searched for diverse activities, functions and influences (1), retrieving relations from the literature (2).

Research in natural language processing (NLP) has provided important resources to extract *fine-grained* relations from scientific papers in specific areas, such as certain subfields of biomedicine (Kim et al., 2013; Nye et al., 2018) or computer science (Wadden et al., 2019). However, these cover only a fraction of all concepts in the literature; in biomedicine alone, there are myriad concepts (Salvadores et al., 2013) not covered by NLP resources. For COVID-19 research, the challenge is especially pronounced due to diversity and emerging concepts; even reading just one paper may require background knowledge in multiple biomedical subfields, physics, chemistry, engineering, computer science and the social sciences. For example,

[*] Equal contribution.
[1] https://covidmechanisms.apps.allenai.org/

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4489–4503
June 6–11, 2021.

consider a paper studying the indoor dynamics of aerosolized SARS-CoV-2 and the effect of ventilation on transmission by using simulation models, or work on economic impacts of COVID-19 on prices and consumption.

To make progress in consolidating such diverse information, we introduce a unified schema of *mechanisms* as a *unified language* covering activities, functions and influences across the sciences. These can be proteins that block viral binding, algorithms to design drugs, the effect heat has on viruses, or COVID-19 has on food prices (Fig. 1).

We build on the fact that mechanisms underlie much of the natural language of scientific papers (Röhl, 2012), and construct a unified schema with two coarse-grained mechanism relations:

- *Direct Mechanisms*: mechanistic *activities* (e.g., viral binding) or *functions* engendered by natural or artificial entities (e.g., a protein used for binding or algorithm used for diagnosis).
- *Indirect Mechanisms*: *influences and associations* such as economic effects of COVID-19 or complications associated with medical procedures.

Our coarse-grained relation schema, over free-form text spans, strikes a balance between the granular information extracted by Closed-IE approaches (Freitag, 1998; Hoffmann et al., 2010) and the schema-free breadth of Open IE approaches (Etzioni et al., 2008; Stanovsky et al., 2018), which often lead to generic and uninformative relations for scientific applications (Kruiper et al., 2020).

Furthermore, our schema facilitates construction of a high-quality KB that synthesizes interdisciplinary knowledge. We construct precisely this, releasing MECHANIC (**Mech**anisms **AN**otated in **C**OVID-19 papers) – an annotated dataset of 2,400 mechanisms based on our schema. We train a state-of-the-art model to extract this information from scientific papers, and use it to build COMB (**C**OVID-19 **O**pen **M**echanism Knowledge **B**ase) – a broad-coverage KB of 1.5M mechanisms in COVID-19 papers. We analyze the characteristics of COMB, showing the distribution of relations across scientific subfields and comparing their quality to other IE approaches.

We demonstrate the utility of COMB in two studies with experts. In the first study, our system achieves high precision and recall in scientific search with structured queries on both diverse viral mechanisms and applications of AI in the literature. In the second study, we evaluate COMB in a usability study with MDs active in treating and researching COVID-19. Our system is rated higher than PubMed search by the clinical experts, in terms of utility and quality.

Our main contributions include:

- We introduce a unified schema for *mechanisms* that generalizes across many types of activities, functions and influences. We construct and distribute MECHANIC, an annotated dataset of papers related to COVID-19, with 2,400 instances of our mechanism relation.
- Using MECHANIC, we train an IE model and apply it to 160K abstracts in COVID-19 literature, constructing COMB, a KB of 1.5M mechanism instances. Manual evaluation of relations sampled from our KB shows them to have 88% accuracy. We also find a model trained on our data reaches roughly 80% accuracy on a sample of general biomedical papers from across the PubMed corpus, with no additional training, demonstrating the generalization of our approach.
- We showcase the utility of COMB in structured search for mechanisms in the literature. In a study with MDs working to combat COVID-19, our system is rated higher than PubMed search in terms of utility and quality.

2 Related work

Mechanisms in science The concept of *mechanisms*, also referred to as *functional relations*, is fundamental across the sciences. For example mechanisms are described in biomedical ontologies (Burek et al., 2006; Röhl, 2012; Keeling et al., 2019), engineering (Hirtz et al., 2002), and across science. Mechanisms can be natural (e.g., the mechanism by which amylase in saliva breaks down starch into sugar), artificial (electronic devices), non-physical constructs (algorithms, economic policies), and very often a blend (a pacemaker regulating the beating of a heart through electricity and AI algorithms).

Although seemingly intuitive, exact definitions of mechanisms are subject to debate in the philosophy of science (Röhl, 2012; Keeling et al., 2019). An Oxford dictionary definition of mechanisms refers to *a natural or established process by which something takes place or is brought about*. More intricate definitions discuss "complex systems producing a behavior", "entities and activities productive of regular changes", "a structure performing a function in virtue of its parts and operations", or the

Schema	Entity types	Relations	Example
SciERC	CS methods/tasks (free-form spans)	used-for	Use **GNNs** for **relation extraction.**
SemRep	Clinical (drugs, diseases, anatomy . . .)	causes, affects, treats, inhibits, interacts, used . . .	. . . intratympanic **dexamethasone injections** for patients with intractable **Meniere's disease.**
ChemProt	Chemicals, proteins	direct/indirect regulator, inhibitor, activator . . .	**Captopril** inhibited **MMP-9** expressions in right ventricles.
DDI	Drugs	interacts	**Quinolones** may enhance the effect of **Warfarin.**
GENIA	Proteins, cellular entities	binding, modification, regulation . . .	**BMP-6** induced phosphorylation of **Smad1/5/8.**
PICO	Clinical	Interventions, outcomes	The **bestatin** group achieved **longer remission.**
Ours: MECHANIC	Medicine, epidemiology, genetics, molecular bio., CS, math, ecology, economics . . . (free-form)	direct (activities, functions) / indirect (influences, associations)	· **RL** can be used to learn **mitigation policies in epidemiological models.** · **Histophilus-somni** causes **respiratory, reproductive, cardiac and neuronal diseases in cattle.**

Table 1: Our broad concept of mechanisms covers many relations within existing science-IE schemas. The table shows examples of representative schemas, and the types of entities and relations they capture.

distinction between "correlative property changes" and "activity determining how a correlative change is achieved" (Röhl, 2012).

Abstract definitions can help with generalization across many important types of mechanisms. The schema we propose (Sec. 3) is inspired by such definitions, operationalizing them and making them more concrete, and also simple enough for models and human annotators to identify.

Information extraction from scientific texts There is a large body of literature on extracting information from scientific papers, primarily in the biomedical sphere. This information often corresponds to very *specific* types of mechanisms, as shown in Tab. 1. Examples include ChemProt (Li et al., 2016) with mechanisms of chemical-protein regulation, drug interactions in the DDI dataset (Segura Bedmar et al., 2013), genetic and cellular activities/functions in GENIA (Kim et al., 2013), semantic roles of clinical entities (Kilicoglu et al., 2011), PICO interventions and outcomes (Wallace et al., 2016; Nye et al., 2018), and computer science methods/tasks in SciERC (Luan et al., 2018). Such schemas have been used, for example, to extract genomic KBs (Poon et al., 2014) and automate systematic reviews (Nye et al., 2020). Our schema draws on these approaches, but with a much broader reach across concepts seen in COVID-19 papers (Tab. 1, Fig. 2).

An important area in information extraction focuses on *open* concepts, with prominent approaches being Open IE (Etzioni et al., 2008) and Semantic Role Labeling (SRL; Carreras and Màrquez, 2005), which share similar properties and predictions (Stanovsky et al., 2018). While such methods are intended to be domain independent, they perform significantly worse in the scientific domain (Groth et al., 2018). Kruiper et al. (2020) developed a multi-stage process to post-process Open IE outputs, involving trained models and humans to find a balance between generic and fine-grained clusters of relation arguments and omitting noisy clusters. In contrast, our unified schema enables annotating a dataset of mechanism relations between free-form spans and training IE models to automatically generalize across diverse relation types.

Our schema is also related broadly to the task of training reading comprehension models on procedural texts describing scientific processes (such as short paragraphs written by crowd workers to explain photosynthesis in simple language; Dalvi et al., 2018). Our representation of scientific texts in terms of a graph of causal relations can potentially help infer processes across science.

COVID-19 IE Recent work (Verspoor et al., 2020a) has focused on extracting information from the CORD-19 corpus (Wang et al., 2020b). PICO concepts are extracted and visualized in an exploratory interface in the COVID-SEE system (Verspoor et al., 2020b). In Wang et al. (2020a), genes, diseases, chemicals and organisms are extracted and linked to existing biomedical KBs with information such as gene-disease relations. Additional relations based on the GENIA schema are extracted from the text. To address the novel COVID-19 domain, the schema is enriched with new entity types

such as viral proteins and immune responses.

In this paper, we focus on a more general schema that captures diverse concepts appearing in literature related to COVID-19, an emerging domain with novel concepts coming from many fields and subfields. The mechanism KG we construct includes—as a subset —diverse biomolecular and clinical information (such as chemical-disease relations) as part of a general mechanism schema.

3 Mechanism Relation Schema

We present a schema that builds upon and consolidates many of the types of mechanisms discussed in Sec. 2. Our defined schema has three key properties: (1) it uses a generalized concept of mechanism relations, capturing specific types of mechanisms in existing schema and extending them broadly; (2) it includes flexible, generic entities not limited to predefined types, and (3) it is simple enough for human annotators and models to identify in the natural language of scientific texts. This schema enables forming our KB by identifying a set of mechanism relations in a corpus of scientific documents (Sec. 4.3).

We formally define each mechanism as a relation $(E_1, E_2, \texttt{class})$ between entities E_1 and E_2, where each entity E is a text span and the `class` indicates the type of the mechanism relation. Entities all share a single common type and can be either natural (e.g., protein functions, viral mechanistic activities) or artificial (e.g., algorithms, devices), to capture the generality of the concepts in science (see Fig. 2). We allow each entity to take part in multiple relations (tuples) within a given text, leading to a "mechanism graph". Mechanisms are categorized into two coarse-grained classes:[2]

Direct mechanisms include *activities* of a mechanistic nature – actions explicitly performed by an entity, such as descriptions of a virus binding to a cell, and explicit references to a function (e.g., a use of a drug for treatment, or the use of AI for drug design as in Fig. 1).

Indirect mechanisms include influences or associations without explicit mechanistic information or mention of a function (such as describing observed effects, without the process involved). These relations correspond more to "input-output correlations" (Röhl, 2012), such as indicating that COVID-19 may lead to economic impacts but not *how* (Fig. 1), as opposed to direct mechanisms describing "inner workings" – revealing more of the intermediate states that lead from initial conditions (COVID-19) to final states (price inflation) or explicitly describing a function. As an example for the utility of this distinction between direct and indirect relations, consider an MD looking to generate a structured list of all *uses* of a treatment (direct mechanism), but not include side effects or complications (indirect).

[2]We also provide a dataset and extraction model for ternary relations in the form of *(subject, object, predicate)*. We focus on the coarse-grained mechanism schema due its broader flexibility and coverage. See App. A.1 for details.

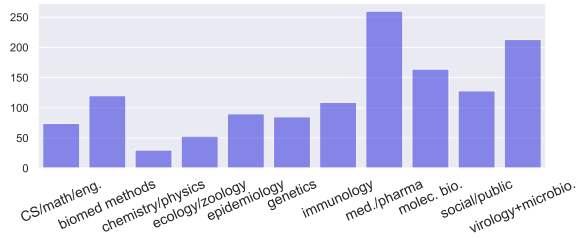


Figure 2: MECHANIC covers a diverse set of scientific fields. Histogram of domains in MECHANIC (sample of 350 relations). Manually labeled relation entities, based on a list of scientific disciplines from Wikipedia.

4 KB Construction

We describe our approach (depicted in Fig. 3) for extracting a knowledge base of mechanisms using our unified schema. We first curate MECHANIC, an annotated dataset of general mechanisms from a small collection of scientific papers (Sec. 4.1). We then train a model on our annotated data to extract mechanism relations from the entire CORD-19 corpus of scientific papers; we use it to build COMB, a knowledge base of mechanisms across the entire CORD-19 corpus of (Sec. 4.2), which supports semantic search for relations (Sec. 4.3).

4.1 Collecting Mechanism Annotations

We construct a dataset of mechanism relations in texts randomly sampled from the CORD-19 corpus (Wang et al., 2020b) that includes scientific papers connected to COVID-19. To circumvent annotation challenges in scientific datasets (Luan et al., 2018) and ensure high-quality annotations, we follow a three-stage process of (1) annotating entities and relations using biomedical experts, (2) unifying span boundaries with an NLP expert, and (3) verifying annotations with a bio-NLP expert. Our annotation process is a relatively low-resource and generalizable approach for a rapid response to the

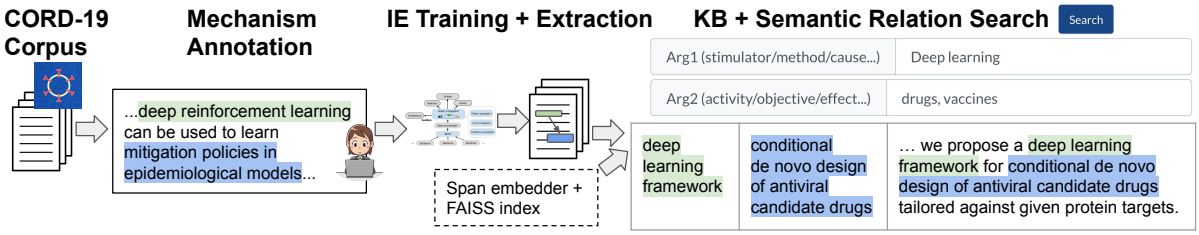


Figure 3: **Overview of our approach**. We collect annotations of mechanisms (textual relations) from the CORD-19 corpus, which are used to train an IE model. We apply the model to over 160K documents in the corpus, extracting over 1.5M relations that are fed into our KB. Entity mention spans are embedded with a language model tuned for semantic similarity, and indexed with FAISS for fast similarity search as part of our search interface.

COVID-19 emergency.

In the first stage, five annotators with biomedical and engineering background annotate all mechanism relations as defined in Sec. 3 (full annotation guidelines are available in our code repository). Relations are annotated as either direct/indirect. Entities are annotated as the longest span of text that is involved in a relation with another entity, while not including redundant or irrelevant tokens. As in related tasks (Luan et al., 2018), annotators are guided to resolve doubt on span boundaries by selecting the longest relevant span.

Annotators had a one-hour training session. In the first part of the training session, annotation guidelines were reviewed. The guidelines included simple explanations of direct/indirect mechanisms along with introductory examples (e.g., "*the virus makes use of spike protein to bind to a cell*", "*A virus leads to respiratory infection*"). In the second part, annotators saw examples from papers in the annotation interface (see Fig. 6, App. A), and performed a few live training annotations.

We initially observed significant variation between annotators in identifying span boundaries for entity annotations, stemming from inherent subjectivity in such annotation tasks (Stanovsky et al., 2018; Luan et al., 2018) and from lack of NLP experience by some annotators. In the second stage, an NLP expert annotator conducted a round of style unification by viewing annotations and adjusting span boundaries to be more cohesive while preserving the original meaning, focusing on boundaries that capture essential but not redundant or generic information (e.g., adjusting the span *substantial **virus replication** by unknown mechanisms* to include only *virus replication*). Finally, in the third stage, a bio-NLP expert with experience in annotating scientific papers verified the annotations and corrected them as needed. The expert accepted 81% of the annotations from the second stage without modification, confirming the high quality of the stage-2 data. Relation label mismatches accounted for 5% of the remaining 19%. Other sources of disagreement were span mismatches and new relations added by the bio-NLP expert adjudicator.

The resulting dataset (MECHANIC: **Mech**anisms **AN**otated in **C**OVID-19 papers) contains 2,370 relation instances (1645 direct, 725 indirect) appearing in 1,000 sentences from 250 abstracts.[3] Average span length is 4 tokens, while the average distance between relation arguments is 11.40 tokens.

4.2 Extracting a KB of Mechanisms

Using MECHANIC, we train an IE model to extract mechanism relations from sentences in scientific documents. We train DyGIE++ (Wadden et al., 2019), a state-of-the-art end-to-end IE model which extracts entities and relations jointly (without assuming to have entity spans given), classifying each relation as one of {DIRECT, INDIRECT}.[4]

To form our corpus-level KB, we apply the trained model to each document in our corpus (all 160K abstracts in the CORD-19 corpus) to extract mechanism relations and then integrate the extracted relations. We find that our trained model achieves high precision scores for high confidence predictions (precision $\geq$ 80% within top-20 predicted relations; see $P@K$ figure, App. B). Therefore, our corpus-level KB is constructed by filtering predictions with low confidence.

[3] The dataset is similar in size to related scientific IE datasets (Luan et al., 2018) which share related challenges in collecting expert annotations of complex or ambiguous concepts over difficult texts.

[4] We use DyGIE++ with SciBERT (Beltagy et al., 2019) embeddings fine-tuned on our task and perform hyperparameter grid search (for dropout and learning rate only) and select the best-performing model on the development set ($7e-4$ and 0.43, respectively). Full details are in App. B.3.

To integrate relations and entities across the corpus, we use standard surface-level string normalization (such as removing punctuation, lemmatizing, and lowercasing) and unify and normalize entity mentions using coreference clusters of entities within a document.[5] Each coreference cluster is assigned a representative entity as the mention with the longest span of text, and all other entities in that cluster are replaced with the representative entity. This is particularly useful for normalizing pronouns such as *it* with the original mention they referred to (e.g., a specific virus or method *it* refers to).

Our final KB (COMB) consists of 1.5M relations in the form of $(E_1, E_2, \texttt{DIRECT/INDIRECT})$ filtered by high confidence score (>= 90%), where entities E_i are standardized free-form spans of text.

4.3 Semantic Relation Search

The constructed KB enables applications for retrieving relations across concepts from many disciplines. For example, searching for all documents that include mechanisms to incorporate *AI* in studies of *heart disease* (E_1 = AI, E_2 = heart disease, $\texttt{DIRECT}$) requires going beyond simply finding documents that mention *AI* and *heart disease*. Here, we describe our approach for searching over the KB by encoding entities and relations, capturing related concepts (such as *cardiac disease* and *heart conditions*), as well as simpler surface matches (*artificial intelligence* ***methods***, *artificial intelligence* ***models***).

Specifically, for a given query $\mathbf{q} := (E_1^q, E_2^q, \texttt{class})$, our goal is to find mechanisms r_i in COMB whose entities are free-form texts similar to E_1^q, E_2^q in the query. The $\texttt{class}$ is used to filter for the type of relation—for example, when explicitly requiring $\texttt{DIRECT}$ mechanisms.

Entity encoding We obtain an encoding function $f : E \mapsto \mathbb{R}^d$ to encode all unique spans (entities) in the KB to a d dimensional vector space. The encoding function is derived by fine-tuning a language model (LM) originally trained on PubMed papers (Gururangan et al., 2020) on semantic similarity tasks. For fine-tuning, we use sentence pairs in STS (Cer et al., 2017) and SNLI (Bowman et al., 2015) following Reimers and Gurevych (2019), and add biomedical sentence pairs from the BIOSSES dataset (Soğancıoğlu et al., 2017).

[5] We use a pre-trained DyGIE++ model trained on SciERC to obtain coreference clusters.

Relation similarity Given a query $\mathbf{q}$, we rank the set of all COMB relations with the same $\texttt{class}$ as the query. For each candidate relation $r = (E_1, E_2, \texttt{class})$ in COMB, we compute its similarity to the query relation $\mathbf{q}$ as the minimum similarity between encodings of their corresponding entities: $\min_{j \in \{1,2\}} f(E_j) \cdot f(E_j^q)$. With this definition, a relation (E_1, E_2) with E_1 very similar to the first entity of the query E_1^q but E_2 distant from E_2^q will be ranked low. For example, with the query $(E_1^q$ = deep learning, E_2^q = drugs$)$, the relation $(E_1$ = microscope, E_2 = drugs$)$ will be ranked low due to the pair (deep learning, microscope). For efficient search, we create an index of embeddings corresponding to the 900K unique surface forms in COMB and employ a system designed for fast similarity-based search (Johnson et al., 2017).

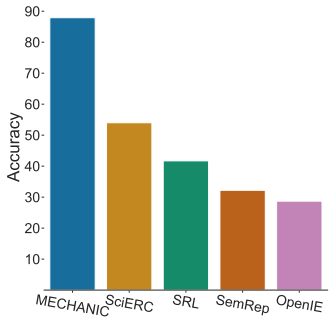
5 Evaluating COMB

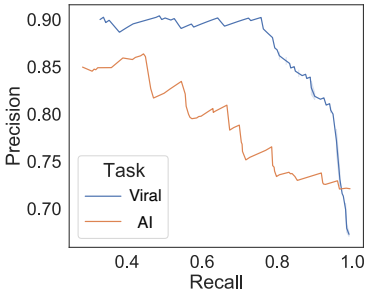
In this section, we evaluate the constructed KB of mechanisms in terms of correctness and informativeness (Sec. 5.1), and its utility in searching for mechanisms (Sec. 5.2). Our main goal is to ensure the mechanism relations have high quality to support our large-scale KB and search applications. We further show that our schema is useful as compared to other schema.

5.1 KB Correctness and Informativeness

We employ two annotators with biomedical and CS backgrounds to judge the quality of the predicted relations in COMB. In particular, following Groth et al. (2018), annotators are given a predicted relation together with the sentence from which it was extracted. We collapse all entities/relations into one generic type for this analysis. Annotators are asked to label the predicted relation as correct if (1) it accurately reflects a mechanistic relation mentioned in the sentence (*correctness*), and (2) the extracted entities and relation label are sufficient to convey the meaning of the relation, without referring to the source sentence (*informativeness*). We collect human judgements for 300 predicted relations for our approach and baselines, sampled from 150 randomly selected sentences. Agreement is 71% by Cohen's Kappa and 73% by Matthew's Correlation Coefficient.

Comparing KB quality to other schemas To showcase the benefit of our approach, we compare the relations extracted using a DyGIE model trained on MECHANIC, versus a DyGIE model





Metric	PubMed	COMB
Search	71%	**90%**
Utility	69.5%	**92%**
Interface	78%	**90%**
Overall	74%	**91%**

Figure 4: Evaluating COMB in studies with experts. COMB is found to have high quality and utility, outperforming other approaches. **Left:** COMB outperforms external resources with either *specific* types of mechanisms or *open* relations, in human evaluation for correctness and usefulness of predictions, on a sample of 300 predicted relations. **Center:** Retrieved relations are ranked by query similarity (Sec. 4.3) and compared to human relevance labels to compute precision/recall. Higher-ranked results are overall judged as relevant by humans. **Right:** Results of COMB search study with five practicing MDs, using both our system and PubMed to search the literature. Experts were given a post-study questionnaire with questions grouped by subject (search, utility, interface). Our mechanism search system performed substantially better than PubMed.

trained on other resources that are most related to our mechanisms: SemRep (Kilicoglu et al., 2011) captures a wide range of biomedical relations (such as drug-drug interactions), and SciERC (Luan et al., 2018) contains relations relevant to computer science (such as "method-task" and "used-for" relations).[6] In addition, we compare with a Semantic Role Labeling (SRL) method (Shi and Lin, 2019) that captures broad relations between free-form spans that focus on agents and actions, and a neural OpenIE model (Stanovsky et al., 2018).

Fig. 4 (left) shows that 88% of relations from COMB are marked as correct by human raters, demonstrating that our approach extracts mechanism relations with better quality than external resources.[7] These results suggest that our predicted relations are of overall high quality and can be used to build our corpus-level KB and explore its utility.

Examining Generalization COVID-19 papers are highly diverse both topically and chronologically. We conduct a small-scale preliminary experiment examining whether a model trained on MECHANIC can generalize to capture mechanism relations in the general biomedical papers, from a much larger corpus of open access papers on PubMed Central (PMC).[8] We randomly sample a set of 200 predicted relations from papers across the entire PMC corpus, and label them using the same criteria used above. As expected, we find that performance drops, but encouragingly is still considerably high: after filtering predictions with confidence lower than 90% in the same way we construct COMB, 76% of relations are considered correct. When filtering for confidence with a threshold of 95% (which captures 70% of the samples), the rate of correct predictions is 78%. In future work it would be interesting to fine-tune our model on a small set of labeled examples from the general PMC corpus to potentially improve these results.

5.2 COMB Utility

We design several search tasks and user studies to evaluate the utility of the constructed KB (Sec. 5.2.1) and compare it with the PubMed medical KB and search engine (Sec. 5.2.2), as judged by medical doctors working on the front lines of COVID-19 treatment and research. All tasks are designed to evaluate our framework's utility in helping researchers and clinicians looking to quickly search for mechanisms or cause-effect relations in the literature and retrieve a list of structured results.

5.2.1 Search Quality

We form search queries based on a wide range of topics pertaining to (1) SARS-CoV-2 mechanisms (such as modes of transmission, drug effects, climatic influences, molecular-level properties) and

[6] We use an expert annotator to align external resources to our `direct` or `indirect mechanism` annotations, e.g., USED-FOR is mapped to *direct* mechanism).

[7] We also experiment with automated evaluation. We split MECHANIC into train/dev/test sets (170/30/50 abstracts), and obtain $F1 = 50.2$ for entity detection, $F1 = 45.6$ for relation detection and $F1 = 42.8$ for classification, on par with performance in other similar scientific IE tasks (Luan et al., 2018). See more details in App. B.4.

[8] https://www.ncbi.nlm.nih.gov/pmc/

Relation query	Example results from KB search interface
E1 Warm climate E2 Coronavirus	Experimental data showed that coronavirus survival was negatively impacted by ozone and high temperature.
E1 COVID-19 E2 Bilateral ground glass opacities	The typical features of COVID-19 in chest CT include bilateral, peripheral, and multifocal ground-glass opacities with or without superimposed consolidations.
E1 Aerosols E2 SARS-CoV-2 transmission	SARS-CoV-2 is transmitted efficiently via the air (via respiratory droplets and/or aerosols) between ferrets.

(a) **Viral mechanism search**. Queries for $(E1, E2)$ relations, and example retrieved results.

E1 Convolutional neural networks E2	3D patch - based convolutional neural networks were trained to predict conductivity maps from B1 transceive phase data.
E1 Computer vision E2	We present a computational pipeline using algorithms from computer vision to decompose ciliary motion into quantitative elemental components.
E1 Graph neural networks E2	We propose a new model called GraphDTA that represents drugs as graphs and uses graph neural networks to predict drug – target affinity.

(b) **AI search**. Queries consists of only E_1, to find all applications of AI approaches/areas.

Table 2: Example search queries and results for the viral mechanism and AI applications tasks.

(2) applications of AI in this area. Tab. 2a and 2b show queries and example relations returned from COMB, along with the context sentences from which they were extracted.

Viral mechanism search Queries are formed based on statements in recent scientific claim-verification work (Wadden et al., 2020; see full list in App. C.2). For example, for the statement *the coronavirus cannot thrive in warmer climates*, we form the query as (E_1 = Warm climate, E_2 = coronavirus) (see Tab. 2a row 1). For statements reflecting an indirect association/influence, we filter for `INDIRECT` relations (Tab. 2a row 2). For statements that reflect an undirected mechanism relation (e.g., *Lymphopenia is associated with severe COVID-19 disease*), we query for both directions.

AI applications search This task is designed to explore the uses of AI in COVID-19 papers (Tab. 2b). We use queries where the first entity E_1 is a leading subfield or method within AI (e.g., *deep reinforcement learning* or *text analysis*), and the second entity E_2 is left unspecified. Since all queries relate to *uses* of AI, we filter for `DIRECT` relations. These open-ended queries simulate an exploratory search scenario, and can potentially surface inspirations for new applications of AI against COVID-19 or help users discover where AI is being harnessed.

Evaluation Expert annotators are instructed to judge if a relation is related to the query or not and if the sentence actually expresses the mechanism. These annotations are used as ground-truth labels to compute precision/recall scores of the relations extracted by our algorithm. Since it is not feasible to label every relation, annotators are shown a list of 20 relations for each query including high and low rank relations returned by our search algorithm.[9] In total, we use 5 annotators to obtain 1,700 relevance labels across both tasks. Inter-annotator agreement is high by several metrics, ranging from 0.7–0.8 depending on the metric and task; see App. C.2. Annotators have graduate/PhD-level background in medicine or biology (for the first task) and CS or biology (for the second task).

Results Fig. 4 (center) shows our results for both tasks. For biomedical search queries, we observe 90% precision that remains stable for recall values as high as 70%. For *AI applications* we observe a precision of 85% at a recall of 40% that drops more quickly. This lower precision is likely due to the fact that E_2 is unspecified, leading to a wider range of results with more variable quality.

Overall, these results showcase the effectiveness of our approach in searching for mechanisms be-

[9] Specifically, for each query we retrieve the top-1000 similar relations from COMB, ranked as described in Sec. 4, and select the top and bottom 10 relations (20 per query, 200(=20x10) per task, 400(=200x2) in total), shuffle their order, and present to annotators together with the original sentence from which each relation was extracted.

tween diverse concepts in COVID-19 papers.

5.2.2 Comparing COMB with PubMed

This experiment compares the utility of COMB in structured search for causal relationships of clinical relevance to COVID-19 with PubMed[10]—a prominent search engine for biomedical literature that clinicians and researchers frequently peruse as their go-to tool. PubMed allows users to control structure (e.g., with MeSH terms or pharmacological actions), is supported by a KB of biomedical entities used for automatic query expansion, and has many other functions.

Expert evaluation We recruit five expert MDs—with a wide range of specialities including gastroenterology, cardiology, pulmonary and critical care—who are active in treating COVID-19 patients and in research. Each expert completed search randomly ordered tasks using both PubMed and our COMB UI, showing the full set of ranked relations, as well as the sentence snippet mentioning the relation, the paper title, and hyperlink to abstract. At the end of the study after all search tasks are completed for both our system and PubMed, experts are given a questionnaire of 21 7-point Likert-scale questions to judge system utility, interface, and search quality. The first 16 questions are taken from a Post Study System Usability Questionnaire (PSSUQ; Lewis, 2002) widely used in system quality research. The last 5 questions are designed by the authors to evaluate search quality such as overall result relevance and ranking (for the full question list, see App. C.2). Each question is asked twice, once for PubMed and once for our system, leading to 21×2×5 = 210 responses.

Search queries We provide experts with seven search queries that were created by an expert medical researcher, relating to causal links (e.g., between COVID-19 and cardiac arrhythmias) and functions (e.g., Ivermectin as a treatment). See full set of queries in App. C.

Results Fig. 4 (right) shows the average Likert scores (normalized to [0%,100%]) across all questions and users for COMB and PubMed. The results show that the medical experts strongly prefer COMB to PubMed (overall average of 91% vs. 74%, with non-normalized scores of 6.6 vs. 5.2). On average across the 21 questions, the majority of the five experts assigned our interface a higher score than PubMed, at an average rate of 3.5/5. This rate increases further when considering ties—on average 4.75/5 of the experts assigned our system a score equal or higher than PubMed.

Overall, our system significantly outperforms PubMed in this task, with an average gap of roughly 20% for search and utility-related questions (Wilcoxon signed rank test p-value is significant at 4.77×10^{-7}). These results are particularly interesting and indicate the potential of COMB because of the experts' strong familiarity with PubMed and the simple nature of our UI.

Our system searches and retrieves *relations*—only texts explicitly mentioning relations that match the input query. This often more precisely reflects the query than results returned by PubMed, which do not have the additional layer of structured information in COMB. For example, for the query (E_1=cardiac arrhythmias, E_2=COVID-19), PubMed returns the following title of one paper: *Guidance for cardiac electrophysiology during the* ***COVID-19*** *pandemic [....] Electrocardiography and* ***Arrhythmias*** *Committee*—E_1 and E_2 are both mentioned, but not within a mechanism relation.

6 Conclusion

We introduced a unified schema for *mechanisms* that generalizes across many types of activities, functions and influences. We constructed and distributed MECHANIC, a dataset of papers related to COVID-19 annotated with this schema. We trained an IE model and applied it to COVID-19 literature, constructing COMB, a KB of 1.5M mechanisms. We showcased the utility of COMB in structured search for mechanism relations in COVID-19 literature. In a study with MDs active in the fight against the disease, our system is rated higher than PubMed search for both utility and quality. Our unified view of mechanisms can help generalize and scale the study of COVID-19 and related areas. More broadly, we envision a KB of mechanisms that enables the transfer of ideas across the literature (Hope et al., 2017), such as by finding relationships between mechanisms in SARS-CoV-2 and other viruses, and assists in literature-based discovery (Swanson and Smalheiser, 1996) by finding cross-document causal links.

Ethical considerations

Our knowledge-base and search system is primarily intended to be used by biomedical researchers working on COVID-19, and researchers from more

[10] https://pubmed.ncbi.nlm.nih.gov/

general areas across science. Models trained and developed on our dataset are likely to serve researchers working on COVID-19 information extraction, and scientific NLP more broadly. We hope our system will be helpful for accelerating the pace of scientific discovery, in the race against COVID-19 and beyond.

Our knowledge-base can include incorrect information to the extent that scientific papers can have wrong information. Our KB includes metadata on the original paper from which the information was extracted, such as journal/venue and URL. Our KB can also miss information included in some papers.

Our data collection process respected intellectual property, using abstracts from CORD-19 (Wang et al., 2020b), an open collection of COVID-19 papers. Our knowledge-base fully attributes all information to the original papers. All annotators were given extensive background on our objectives, and told their annotations will help build and evaluate a knowledge-base and search engine over COVID-19 research. Graduate-student annotators were payed 25 USD per hour. MD experts helped evaluate the tool on a voluntary basis.

Acknowledgements

We like to acknowledge a grant from ONR N00014-18-1-2826. Authors would also like to thank anonymous reviewers, members of AI2, UW-NLP and the H2Lab at The University of Washington for their valuable feedback and comments.

References

Matt Apuzzo and David D. Kirkpatrick. 2020. Covid-19 changed how the world does science, together. *New York Times*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *EMNLP*.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*.

Patryk Burek, Robert Hoehndorf, Frank Loebe, Johann Visagie, Heinrich Herre, and Janet Kelso. 2006. A top-level ontology of functions and its application in the open biomedical ontologies. *Bioinformatics*.

Xavier Carreras and Lluís Màrquez. 2005. Introduction to the conll-2005 shared task: Semantic role labeling. In *CoNLL)*.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *SemEval-2017*.

Bhavana Dalvi, Lifu Huang, Niket Tandon, Wen-tau Yih, and Peter Clark. 2018. Tracking state changes in procedural text: a challenge dataset and models for process paragraph comprehension. In *NAACL*.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. In *EMNLP*.

Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S Weld. 2008. Open information extraction from the web. *Communications of the ACM*.

D. Freitag. 1998. Toward general-purpose learning for information extraction. In *COLING-ACL*.

Paul Groth, Mike Lauruhn, Antony Scerri, and Ron Daniel Jr. 2018. Open information extraction on scientific text: An evaluation. In *CoNLL*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *ACL*.

Julie Hirtz, Robert B Stone, Daniel A McAdams, Simon Szykman, and Kristin L Wood. 2002. A functional basis for engineering design: reconciling and evolving previous efforts. *Research in engineering Design*.

R. Hoffmann, Congle Zhang, and Daniel S. Weld. 2010. Learning 5000 relational extractors. In *ACL*.

Tom Hope, Joel Chan, Aniket Kittur, and Dafna Shahaf. 2017. Accelerating innovation through analogy mining. In *KDD*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*.

Diane Marie Keeling, Patricia Garza, Charisse Michelle Nartey, and Anne-Ruxandra Carvunis. 2019. Philosophy of biology: The meanings of'function'in biology and the problematic case of de novo gene emergence. *Elife*, 8:e47014.

Halil Kilicoglu, Graciela Rosemblat, Marcelo Fiszman, and Thomas C Rindflesch. 2011. Constructing a semantic predication gold standard from the biomedical literature. *BMC bioinformatics*.

Jin-Dong Kim, Yue Wang, and Yamamoto Yasunori. 2013. The genia event extraction shared task, 2013 edition-overview. In *BioNLP Shared Task Workshop*.

Ruben Kruiper, Julian Vincent, Jessica Chen-Burger, Marc Desmulliez, and Ioannis Konstas. 2020. In layman's terms: Semi-open relation extraction from scientific texts. In *ACL*.

James R Lewis. 2002. Psychometric evaluation of the pssuq using data from five years of usability studies. *International Journal of Human-Computer Interaction*.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*.

I. Loshchilov and F. Hutter. 2017. Fixing weight decay regularization in adam. *ArXiv*, abs/1711.05101.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *EMNLP*.

Ines Montani and Matthew Honnibal. 2018. Prodigy: A new annotation tool for radically efficient machine teaching. *Artificial Intelligence*.

Lidia Morawska, Donald K Milton, et al. 2020. It is time to address airborne transmission of covid-19.

Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain J Marshall, Ani Nenkova, and Byron C Wallace. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *ACL*.

Benjamin E Nye, Jay DeYoung, Eric Lehman, Ani Nenkova, Iain J Marshall, and Byron C Wallace. 2020. Understanding clinical trial reports: Extracting medical entities and their relations. *arXiv preprint arXiv:2010.03550*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Hoifung Poon, Chris Quirk, Charlie DeZiel, and David Heckerman. 2014. Literome: Pubmed-scale genomic knowledge base in the cloud. *Bioinformatics*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP*.

Johannes Röhl. 2012. Mechanisms in biomedical ontology. In *Journal of Biomedical Semantics*. BioMed Central.

Manuel Salvadores, Paul R Alexander, Mark A Musen, and Natalya F Noy. 2013. Bioportal as a dataset of linked biomedical ontologies and terminologies in rdf. *Semantic web*.

Isabel Segura Bedmar, Paloma Martínez, and María Herrero Zazo. 2013. Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). ACL.

Peng Shi and Jimmy Lin. 2019. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*.

Gizem Soğancıoğlu, Hakime Öztürk, and Arzucan Özgür. 2017. Biosses: a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics*.

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. Supervised open information extraction. In *NAACL*.

D. Swanson and N. Smalheiser. 1996. Undiscovered public knowledge: A ten-year update. In *KDD*.

Karin Verspoor, K Bretonnel Cohen, Michael Conway, Berry de Bruijn, Mark Dredze, Rada Mihalcea, and Byron C Wallace. 2020a. Proceedings of the 1st workshop on nlp for covid-19 (part 2) at emnlp 2020. In *Workshop on NLP for COVID-19 at EMNLP 2020*.

Karin Verspoor, Simon Šuster, Yulia Otmakhova, Shevon Mendis, Zenan Zhai, Biaoyan Fang, Jey Han Lau, Timothy Baldwin, Antonio Jimeno Yepes, and David Martinez. 2020b. Covid-see: Scientific evidence explorer for covid-19 related research. *arXiv preprint arXiv:2008.07880*.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *EMNLP*.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. In *EMNLP*.

Byron C Wallace, Joël Kuiper, Aakash Sharma, Mingxi Zhu, and Iain J Marshall. 2016. Extracting pico sentences from clinical trial reports using supervised distant supervision. *The Journal of Machine Learning Research*.

Jingqi Wang, Huy Anh Pham, Frank Manion, Masoud Rouhizadeh, and Yaoyun Zhang. 2020a. Covid-19 signsym: A fast adaptation of general clinical nlp tools to identify and normalize covid-19 signs and symptoms to omop common data model. *arXiv preprint arXiv:2007.10286*.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, et al. 2020b. Cord-19: The covid-19 open research dataset. *arXiv preprint arXiv:2004.10706*.

A Data Annotation

A.1 Granular Relations

In addition to the two coarse-grained relation classes, we also experimented with *granular* relations where the `class` represents a specific type of a mechanism relation explicitly mentioned in the text (we constrain the mention a single token for simplicity, e.g., *binds, causes, reduces*; see Fig. 5 for examples of granular relations). While more granular, these relations are also less general – as the natural language of scientific papers describing mechanisms often does not conform to this more rigid structure (e.g., long-range and implicit causal relations). We thus focus most of our work on coarse-grained relations. We release our dataset and a model for extraction of granular relations to support future research and applications, in our code repository.

Sentences	Granular relations
Viral infections probably **initiate** a large percentage of childhood and adult asthmatic attacks based on a history of preceding ' cold '.	(Viral infections, **initiate,** childhood and adult asthmatic attacks)
PUVA is useful to **treat** human platelet (PTL) concentrates in order to **eliminate** Leishmania spp.	(PUVA, **treat,** human platelet (PTL) concentrates) (PUVA, **eliminate,** Leishmania spp)

Figure 5: Examples of granular relations.

A.2 Annotation Collection

We utilize the Prodigy (Montani and Honnibal, 2018) annotation platform which provides the ability to select span boundaries and relations with ease. Each annotator undergoes a training session in which we cover the definitions of spans and relations as well as use of the platform. See annotation guidelines in our code repository for more details and examples.

Tab. 3 shows examples of differences between annotations, with disagreements in the span boundaries. This reflects the challenging nature of our task with relations between flexible, open entities.

B IE Evaluations

B.1 Automated evaluation metrics

Entity detection Given a boolean span matching function $\mathrm{m}(s_1, s_2) = \mathbb{1}(s_1 \text{ matches } s_2)$, a predicted entity mention $\hat{e}$ is correctly *identified* if there exists some gold mention e^* in $\mathcal{D}$ such that $\mathrm{m}(\hat{e}, e^*) = 1$ (since there is only one entity type, an entity mention is correctly classified as long as its span is correctly identified).

Following common practice in work on Open IE (Stanovsky et al., 2018), we report results using a partial-matching similarity function, in this case based on the widely-used Rouge score: $\mathrm{m}_{\text{rouge}}(s_1, s_2)$ is true if Rouge-L$(s_1, s_2) > 0.5$ (Lin, 2004).

Relation detection / classification Given a boolean span matching function, a predicted coarse-grained relation $\hat{r} = (\hat{E}_1, \hat{E}_2, \hat{y})$ is correctly *identified* if there exists some gold relation $r^* = (E_1^*, E_2^*, y^*)$ in $\mathcal{D}$ such that $\mathrm{m}(\hat{E}_1, E_1^*) = 1$ and $\mathrm{m}(\hat{E}_2, E_2^*) = 1$. It is properly *classified* if, in addition, $\hat{y} = y^*$.

Relation identification measures the model's ability to identify mechanisms of any type - direct or indirect - while relation classification aims to discriminate between direct and indirect types of mechanism mentions in the text.

B.2 Baselines

SemRep The SemRep dataset (Kilicoglu et al., 2011), consisting of 500 sentences from MEDLINE abstracts and annotated for semantic predication. Concepts and relations in this dataset relate to clinical medicine from the UMLS biomedical ontology (Bodenreider, 2004), with entities such as drugs and diseases. Some of the relations correspond to mechanisms (such as X TREATS Y or X CAUSES Y); By the lead of domain experts, we map these existing relations to our mechanism classes and use them to train DyGIE. Other relations are even broader, such as PART-OF or IS-A – we do not attempt to capture these categories as they often do not reflect a functional relation.

Scierc SciERC dataset (Luan et al., 2018), consisting of 500 abstracts from computer science pa-

Context	Annotator 1	Annotator 2
Predicted siRNAs should effectively silence the genes of SARS - CoV-2 during siRNA mediated treatment.	(predicted siRNAs, silence the genes of SARS - CoV-2, DIRECT)	(siRNAs, silence the genes of SARS - CoV-2 during siRNA mediated treatment, DIRECT)
Recent reports show that the inhibition of **NSP4** expression by small interfering RNAs leads to alteration of the production and distribution of other viral proteins and mRNA synthesis , suggesting that NSP4 also **affects virus replication** by unknown mechanisms.	(NSP4, affects virus replication, INDIRECT)	(NSP4, virus replication by unknown mechanisms, INDIRECT)

Table 3: Examples of differences between two annotators. The core meaning of the relation is equivalent across both annotators.

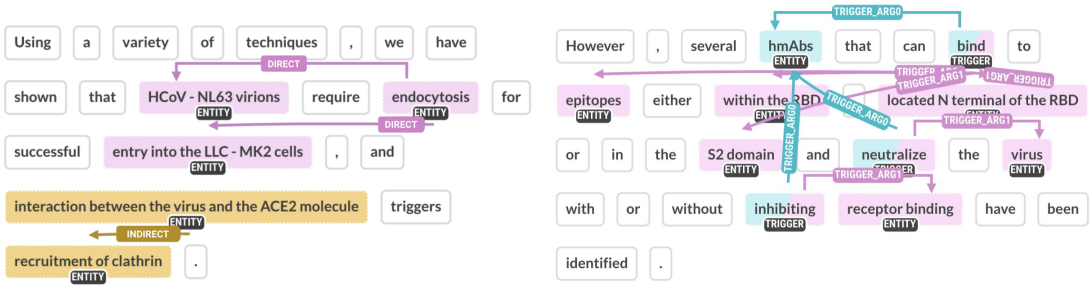


Figure 6: Example of the annotation interface for coarse (left) and granular (right) mechanism relations.

pers that are annotated for a set of relations, including for USED-FOR relations between methods and tasks. We naturally map this relation to our `DIRECT` label and discard other relation types, and use this dataset to train DYGIE.

SRL Finally we also use a recent BERT-based SRL model (Shi and Lin, 2019). We select relations of the form (`Arg0, verb, Arg1`), and evaluate using our partial metrics applied to `Arg0` and `Arg1` respectively.

B.3 Hyperparameter Search

We perform hyperparameter search over these sets of parameters:

- **Dropout** is randomly selected from intervals $[0, 0.5]$.
- **Learning rate** is randomly selected between $[1e-5, 1e-2]$
- **Hidden Size** is randomly selected from interval $[64, 512]$

Hyperparameter search is implemented using grid search with the Allentune library (Dodge et al., 2019). For each experiment we set the search space to be among 30 total samples in hyperparameter space. We select the best-performing parameters using the development set.

B.4 Best Performing Model over MECHANIC

We use the DYGIE package (Wadden et al., 2019) to train models for entity and relation extraction over MECHANIC and we utilize SciBERT (Beltagy et al., 2019) for our text embeddings and finetune upon that, with learning rate for finetuning set to $5e-5$ with weight decay of 0.01. The training was run for 100 epochs with the $slanted\_triangular$ (Howard and Ruder, 2018) learning rate scheduler. We used the AdamW (Loshchilov and Hutter, 2017) optimization algorithm. In our objective function we assign equal weights to relation and span loss terms. The maximum allowed length of spans is 12.

The hyperparameters achieving best performance over our development search are 0.43, $7e-4$ and 215 for dropout, learning rate and hidden size respectively. All other parameters are kept to default values (available in our code repository).

Tab. 4 compares the performance of our best model with the baselines introduced in Sec. B.2. Fig. 7 shows Precision@K results, with our model reaching high absolute numbers.

Model	RC	RD	ED
OpenIE	-	15.5	25.6
SRL	-	24.5	27.7
DYGIE(SemRep)	6.8	8.3	32.5
DYGIE(SciERC)	18.6	20.4	39.2
DYGIE(MECHANIC)	**42.8**	**45.6**	**50.2**

Table 4: F1 scores. Relations from SRL and OpenIE do not map directly to `DIRECT MECHANISM` and `INDIRECT MECHANISM` classes, and do not have relation classification scores.

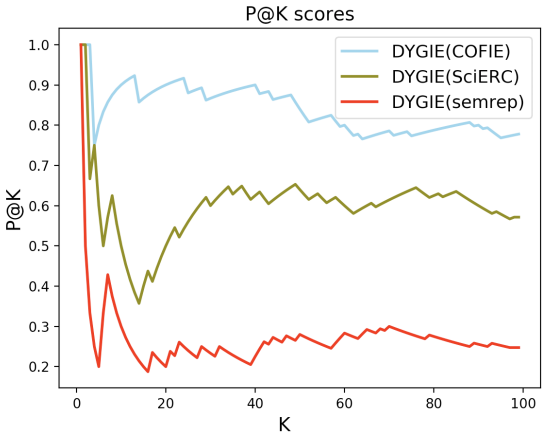


Figure 7: Precision@K of our model compared with pre-trained SciERC and SemRep baselines. P@K for our model is high in absolute numbers.

B.5 Granular relation prediction

Granular relations are evaluated in the same fashion as coarse-grained relations, with the additional requirement that the predicted predicate token $\hat{p}$ must match the gold p^*. Our evaluation shows that the model trained to predict granular triples achieves $F1$ score of 44.0. When predicting relations without trigger labels (i.e., `(s, o)`), the model achieves $F1$ scores of 53.4. These results are not comparable to those for MECHANIC, which includes more documents and relations that did not directly conform to the `(s, o, p)` schema.

B.6 Best Performing Model over MECHANIC-G

Here too we use the DYGIE package (Wadden et al., 2019) with SciBERT (Beltagy et al., 2019). Due to technical equivalence in the annotation schema of our granular relations and the event extraction task in Wadden et al. (2019), we make use of the event extraction functionality of DYGIE. For fine-tuning the embedding weights of SciBERT we used the same learning rate weight as for MECHANIC, and the best hyperparameters found are 0.30, $11e-4$ and 372 for dropout, learning rate and hidden layer size respectively. All other parameters are kept to default values (available in our code repository).

C Human evaluation guidelines

C.1 KB Correctness and Informativeness evaluation guideline

Relation quality evaluations over various domains For the task involving the exploration of viral mechanisms, we used 10 recent scientific claims taken from (Wadden et al., 2020). These 10 claims, and the queries constructed for them, are as follows:

- Remdesevir has exhibited favorable clinical responses when used as a treatment for coronavirus. X = [Remdesevir], Y = [SARS-CoV-2, coronavirus, COVID-19]
- Lopinavir / ritonavir have exhibited favorable clinical responses when used as a treatment for coronavirus. X = [Lopinavir, Ritonavir], Y = [SARS-CoV-2, coronavirus, COVID-19]
- Aerosolized SARS-CoV-2 viral particles can travel further than 6 feet. X = [Air, Aerosols, Droplets, Particles, Distance], Y = [SARS-CoV-2 transmission]
- Chloroquine has shown antiviral efficacy against SARS-CoV-2 in vitro through interference with the ACE2-receptor mediated endocytosis. X = [Chloroquine], Y = [ACE2-receptor, Endocytosis, interference with the ACE2-receptor mediated endocytosis.]
- Lymphopenia is associated with severe COVID-19 disease. X = [Lymphopenia], Y = [severe COVID-19 disease, COVID-19]
- Bilateral ground glass opacities are often seen on chest imaging in COVID-19 patients. X = [Bilateral ground glass opacities], Y = [chest imaging in COVID-19 patients]
- Cardiac injury is common in critical cases of COVID-19. X = [COVID-19], Y = [Cardiac injury]
- Cats are carriers of SARS-CoV-2. X = [Cats], Y = [SARS-CoV-2]

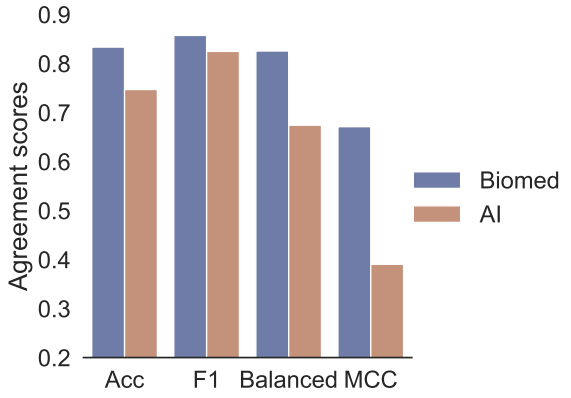


Figure 8: Average pairwise annotator agreement by several metrics. In the AI task human labels were more diverse but with overall high precision / recall.

- Diabetes is a common comorbidity seen in COVID-19 patients. X = [Diabetes], Y = [COVID-19]

- The coronavirus cannot thrive in warmer climates. X = [warmer climates], Y = [coronavirus]

- SARS-CoV-2 binds ACE2 receptor to gain entry into cells. X = [SARS-CoV-2], Y = [binds ACE2 receptor, binds ACE2 receptor to gain entry into cells]

For the AI open-ended search task, we used the following approaches/areas as queries (see guidelines and examples in our code repository): *artificial intelligence, machine learning, statistical models, predictive models, Graph Neural Network model, Convolutional Neural Network model, Recurrent Neural Network model, reinforcement learning, image analysis, text analysis, speech analysis*.

For both tasks, we use the following metrics to measure pairwise agreement between annotators (Fig. 8): standard accuracy (proportion of matching rating labels), F1 (taking into account both precision and recall symmetrically), balanced accuracy (with class weights to down-weight the higher proportion of positive ratings), and finally the Matthew Correlation Coefficient (MCC) score, using the corresponding functions in the Scikit-Learn Python library (Pedregosa et al., 2011).

Comparing KB quality to other schema We sampled the relations predicted by our model and the baseline models introduced in App. B.2. We randomly selected 20 abstracts from the MECHANIC test set and show at most two predictions (if available) for each sentence within that abstract. In total 300 relations are extracted. Each relation is shown separately to two bio-NLP expert annotators (with annotators blind to the condition), who label each relation with a 0/1 label (1 if the relation is both *correct* and *informative*).

Topic	Question
IR	The overall search results accurately matched the query.
IR	I was satisfied by the overall ranking of results.
IR	I found results that were both relevant and interesting or new to me, making this system useful for rapid explorations.
IR	I wanted to read the papers shown.
overall	**Overall, I am satisfied with this system.**
UI/UX	Overall, I am satisfied with how easy it is to use this system.
UI/UX	It was simple to use this system.
UI/UX	I felt comfortable using this system.
UI/UX	The information (such as on-screen messages and other documentation) provided with this system was clear.
UI/UX	The organization of information on the system screens was clear.
UI/UX	The system gave error messages that clearly told me how to fix problems.
UI/UX	Whenever I made a mistake using the system, I could recover easily and quickly.
UI/UX	The interface of this system was pleasant.
UI/UX	I liked using the interface of this system.
UI/UX	It was easy to learn to use this system.
utility	I was able to understand and judge each individual result quickly and without effort.
utility	I was able to complete the tasks and scenarios quickly using this system.
utility	The information was effective in helping me complete the tasks and scenarios.
utility	I believe I could become productive quickly using this system.
utility	It was easy to find the information I needed.
utility	This system has all the functions and capabilities I expect it to have.

Figure 9: List of post-study questions given to MDs.

C.2 KB Utility

MDs are instructed to search with our interface and with PubMed search, with the following 7 topics:

- Query 1: Cardiac arrhythmias caused by COVID 19

- Query 2: Hydroxychloroquine and its effect on COVID 19

- Query 3: Ivermectin and its role in management of COVID 19

- Query 4: Pulmonary embolism effect on complications related to COVID 19

- Query 5: Liver disease and COVID 19

- Query 6 : Inflammatory bowel disease and COVID -19

- Query 7 : Antibody therapy and its uses/effects on COVID-19

The full list of the post-study evaluation questions given to MDs is shown in Fig. 9.